同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第七讲 空间数据：空间自回归

沈煜 博士 副教授
嘉定校区交通运输工程学院311室
yshen@tongji.edu.cn
2022年04月08日

计划进度

周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	
2	2022.03.04	沈煜	在线数据采集方法	爬虫
3	2022.03.11	沈煜	线性回归模型	回归分析
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	广义线性回归（作业1）	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法（作业2）	
8	2022.04.15	沈煜	关联：Apriori	机器学习
9	2022.04.22	沈煜	决策树、支持向量机（作业3）	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络（期末大作业）	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类：K-Means、DBSCAN	
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报（1）	
17	2022.06.17	沈煜	期末汇报（2）	

主要内容

- 空间分析的基本概念
- 空间自相关
  - 空间自相关性统计
  - 空间权重矩阵
- **空间回归模型**
  - **空间滞后模型**
  - **空间误差模型**
  - **模型选择**
- **可视化**

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

空间回归模型

线性回归

- 在位置n，揭示因变量y_n与一系列解释变量x_{nj}间的线性关系：
  - $y_n = \sum_j x_{nj}\beta_j + e_n$
  - 其中，$E[e_n] = 0$，即系统性误差为0.
- 用矩阵表示为：
  - $y_{n\times 1} = X_{n\times j}\beta_{j\times 1} + e_{n\times 1}$
  - 其中，$E[e_{n\times 1}] = 0$.
- 用条件期望表示，即当我们知道X时，y的值平均是多少：
  - $E[y|X] = E[X\beta|X] + E[e|X] = X\beta + 0.$

线性回归

- 普通最小二乘法（OLS）估计参数：
  - $\beta' = (X^T X)^{-1} X^T y$
  - $E[\beta'] = E[(X^T X)^{-1} X^T (X\beta)\ ] + E[(X^T X)^{-1} X^T e\ ]$
  - $E[\beta'] = \beta + 0 = \beta$
- 边际效应：
  - 当X变化时，y的变化情况
  - $E[y|\Delta X] = \Delta X\beta$
  - 在线性回归中，边际效应就是回归的参数

空间计量经济学

- Spatial Econometrics: a subset of econometric methods concerned with spatial aspects present in cross-sectional and space-time observations (Anselin 2006)
- 用于处理横断面或时空观测数据的一系列考虑空间方面特征的计量经济学方法。
- 在模型的建立、评估、诊断与预测中特别需要考虑到位置、距离与排列等因素。
- 空间影响：
  - 空间依赖性（Spatial dependence）
  - 空间异质性（Spatial heterogeneity）

空间影响

- 空间异质性
  - 空间结构变化引起参数变化
- 空间依赖性
  - 二维多方向依赖性
  - 并非时间序列方法的直接扩展

空间计量经济学要解决的问题

- 指明空间依赖性与异质性的结构
- 检验是否存在空间影响
- 建立并估计带有空间影响的模型
- 空间上的预测

空间依赖性

- 在时间序列分析中，y_{t-k}表示将y_t前移t个时间单位
- 在空间分析中，假设一个网格结构，$y_{i-l,j}$，$y_{i+l,j}$，$y_{i,j-l}$，$y_{i,j+l}$分别表示将$y_{i,j}$向北、南、东、西移动l个空间单位

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$

空间滞后（Spatial Lag）

- 相邻值的加权平均
- 相邻（邻居）有空间权重决定
- $y_{iL} = w_{i1}y_1 + w_{i2}y_2 + \cdots + w_{iN}y_N = \sum_j w_{ij}y_j$
- 用矩阵表示：$y_L = Wy$
- Wy：空间滞后因变量

空间滞后与移动平均

- 和移动平均类似，空间滞后是对数据的平滑
- Wy相比原始的y，方差更小
- 但是，空间滞后不是移动平均，因为$w_{ii} = 0$
- 并没有计算移动窗口的“中心点”

考虑空间滞后的回归模型

- 因变量： Wy
  - 空间（自回归）滞后模型
- 自变量：WX
  - 空间交叉回归模型
- 误差项：We
  - 空间（自回归）误差模型

空间滞后模型（Spatial Lag）

- 解析模型中的空间交互关系：空间依赖程度
  - 如peer-effects
  - 体现空间上进行交互影响与反应的过程
- 模型形式
  - $y = \rho W y + X\beta + e$
  - Wy：空间自回归变量
  - ρ：空间自回归参数
  - X：回归自变量

过滤（Spatial Filter）

- 移除空间自相关的影响
- $y = \rho W y + X\beta + e$
- $y - \rho W y = X\beta + e$
- $(I - \rho W)y = X\beta + e$
- $I - \rho W$：空间过滤
- 仍然需要估计ρ值

空间乘子（Spatial Multiplier）

- $(I-\rho W)y=X\beta+e$
- $E(y|\Delta X)=(I-\rho W)^{-1}(\Delta X)\beta$
- $E(y|\Delta X)=[I+\rho W+\rho^2W^2+\cdots](\Delta X)\beta$
- *X*变化后对*y*的影响不止是$(\Delta X)\beta$
- 空间乘子的意义
  - 在位置*i*的变量*x*变化产生的影响不仅影响到*i*本身，还会影响到邻近位置，邻近位置的邻近位置等等
  - 通过*X*的改变，计算变化的空间影响

忽略空间滞后关系的后果

- 忽略了实际存在的空间交互的关系
- 忽略了一些变量
- OLS产生变差与不一致性
- 潜在可能会：
  - 参数、符号、标准差、显著性、模型拟合度等都会出现错误

空间误差模型（Spatial Error）

- 动机：
  - 空间上的依赖性体现在无法观察的误差项中
  - 在空间尺度上的观察结果与实际规律不一致
  - 无法对空间关系作出实质上的解释
  - 评估结果更倾向于空间误差模型的形式

非球面误差的方差

- Non-Spherical Error Variance
- 由于存在空间自相关性，误差的协方差不为0
- 非对角元素为非0值
- $E[ee^T] = \Sigma \neq \sigma^2 I$
- 即：
- 空间上的依赖性使得协方差$E[e_i e_j] \neq 0$，其中$i \neq j$

空间误差模型（Spatial Error）

- $y = X\beta + e$，其中$e = \lambda We + \mu$
- 协方差矩阵$\Sigma = \sigma^2[(I - \lambda W)'(I - \lambda W)]^{-1}$
- 逆矩阵$\Sigma^{-1} = (1/\sigma)^2[(I - \lambda W)'(I - \lambda W)]$
- 空间误差模型可改写为
- $y = X\beta + (I - \lambda W)^{-1}\mu$
  - 不受空间乘子（Spatial Multiplier）影响
  - 空间自相关性的影响主要体现在误差的方差
  - 会对空间预测的结果产生影响

误差与异方差

- 矩阵$\Sigma = \sigma^2[(I - \lambda W)'(I - \lambda W)]^{-1}$的方差，即对角线上的值并不是恒定的，它们受到邻居的数量的影响
- 这会引起误差项e的异方差的问题
  - heteroskedasticity
- 即使μ是同方差的
  - homoscedasticity
- 客观上，我们很难通过模型的异方差来判断真正的异方差的情况

忽略空间误差关系的后果

- 模型效果欠佳
- OLS的拟合结果是无偏的（unbiased），但是无效的（inefficient）。
- 潜在可能会：
  - 参数正确
  - 但参数的标准差与显著性，以及模型的拟合度都会出现错误

模型选择

- 标准回归方法中的一些假设可能不符合实际情况
- 因此，我们需要判断标准回归模型是否符合假设
- 从误差项的分析入手

零假设

- 经典线性回归模型：
- $y = X\beta + e$，其中$E[ee^T] = \sigma^2 I$
- 没有共线性，没有缺失的变量
- 误差项的方差恒定且不相关
- 误差项与自变量不相关

可能的违反假设的情况

- 异方差：误差的方差不恒定
- 空间相关的误差项
- 由于缺失空间滞后自变量的误差相关性的影响
- 以上的情况可能同时存在

H_1假设：空间滞后模型

- 模型实际是$y = \rho Wy + X\beta + e$ （1）
- 模型假设为$y = X\beta + e$ （2）
- （2）相当于当（1）的$\rho = 0$时的情况
- （2）为约束模型（ρ被约束为0）
- （1）为非约束模型（ρ可以为任意值）

H_1假设：空间误差模型

- 模型实际是$y = X\beta + e$，其中$e = \lambda We + \mu$ （1）
- 模型假设为$y = X\beta + e$ （2）
- （2）相当于当（1）的$\lambda = 0$时的情况
- （2）为约束模型（λ被约束为0）
- （1）为非约束模型（λ可以为任意值）

假设检验

- H_0：约束模型；H_1：非约束模型
- 对空间滞后模型：
- H_0：$\rho = 0$；H_1：$\rho \neq 0$
- 对空间误差模型：
- H_0：$\lambda = 0$；H_1：$\lambda \neq 0$

两种检验方式

- 扩散检验（diffuse tests）
  - 没有空间关系VS有空间关系，但不确定哪种
- 聚焦检验（focused tests）
  - 确定具体的空间模型

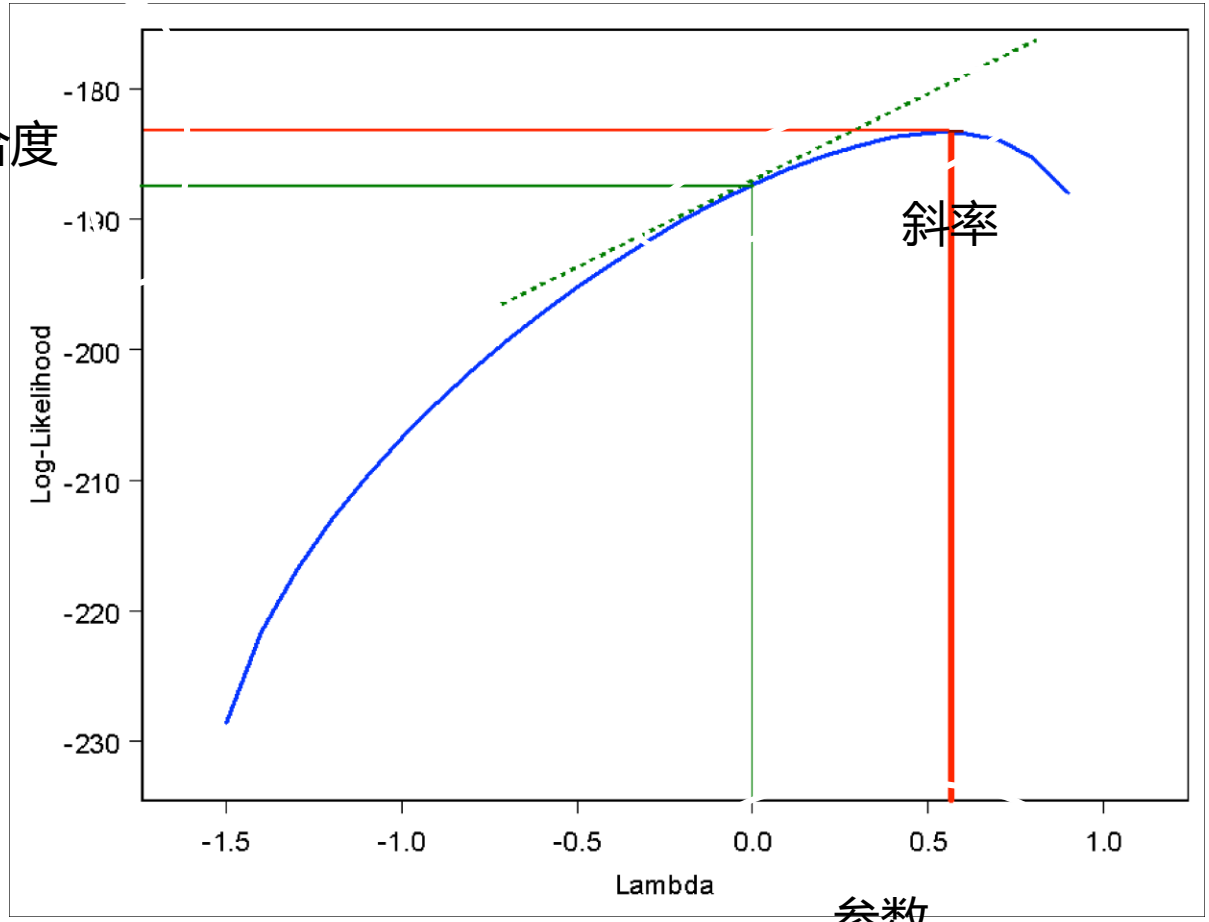
扩散检验

- 拒绝“不存在”空间关系的假设
- 存在空间自相关性，但不具体指明是哪种形式
- 对回归模型的残差（residual）做Moran's *I*检验
  - 残差：$\mu = y - Xb$
  - $I = \frac{\mu^T W \mu / S_0}{\mu^T \mu N}$，$S_0 = \sum_i \sum_j w_{ij}$
  - 对行做标准化，$S_0 = N$，则$I = \mu^T W \mu / \mu^T \mu$
- 无法拒绝零假设证明模型不存在空间上的相关性
- 但是，拒绝零假设并不能指明具体的模型结构
  - 除此之外，还可能是别的情况引起的，如异方差等

聚焦检验

- 方法：
  - 比较受限与非受限两个模型
- 不同的检验策略
  - 比较估计值：将参数估计值与零假设进行比较
  - 比较拟合度：最大似然度与受限模型似然度比较
  - 似然度函数的导数（斜率）：与0比
- 无论是对参数值、拟合度还是斜率的比较，基本思路都是看这些值是否可以拒绝零假设
  - 基于统计方差

参数、拟合度与斜率


拟合度
-180
-190
-200
-210
-220
-230
Log-Likelihood
斜率
-1.5
-1.0
-0.5
0.0
0.5
1.0
Lambda
参数

经典检验方法

- 比较估计值：Wald检验
- 比较拟合度：Likelihood Ratio检验
- 比较斜率：拉格朗日乘子（Lagrange Multiplier）或ρ值（Rao Score）检验

LM误差检验

- (Burridge 1980)
- $\left[\frac{\mu^T W \mu}{\sigma^2}\right]^2 / T \sim \mathrm{X}^2(I)$
- 其中：
  - μ为OLS的残差
  - $T = tr(WW + W^T W)$
  - tr：矩阵的迹，主对角线上各个元素的总和

LM滞后检验

- (Anselin 1988)
- $[e^T Wy/\sigma^2]^2/T_1 \sim \mathrm{X}^2(I)$
- $T_1 = (WXb)^T M(WXb)/\sigma^2 + T$
  - 等式右侧前半部分为WXb在X上的平方的残差的和，即空间滞后的预测值（Xb）的回归
  - $M = I - X(X^T X)^{-1} X^T$

鲁棒（Robust）检验

- 当LM误差和LM滞后检验结果都拒绝另一种模型形式的时候
  - LM误差检验拒绝零假设
  - LM滞后检验拒绝零假设
- 只有当LM误差与LM滞后都拒绝零假设时才考虑鲁棒检验
- 选择最显著的模型形式

模型选择

- 首先建立非受限模型，检测受限模型
  - 需要对较为复杂的空间模型做出估计，检验参数
- 第一步模型选择：LM检验
  - 结果不显著：OLS模型
  - 只有LM误差检验显著：选择空间误差模型
  - 只有LM滞后检验显著：选择空间滞后模型
  - 如果都显著：进行下一步

模型选择

- 第二步模型选择：鲁棒检验
  - 鲁棒LM误差检验显著，鲁棒LM滞后不显著：选择空间误差模型
  - 鲁棒LM滞后检验显著，鲁棒LM误差不显著：选择空间滞后模型
  - 如果两种鲁棒加测模型都显著：选择相对最显著的（即数值最大的）模型

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

可视化：制图原则

地理可视化（Geovisualization）

- 创建或使用可视化元素来辅助思考、理解与知识的构建。
  - The creation and use of visual representations to facilitate thinking, understanding and knowledge construction.
  - 探索、合成、表达、分析

How to Lie with Maps

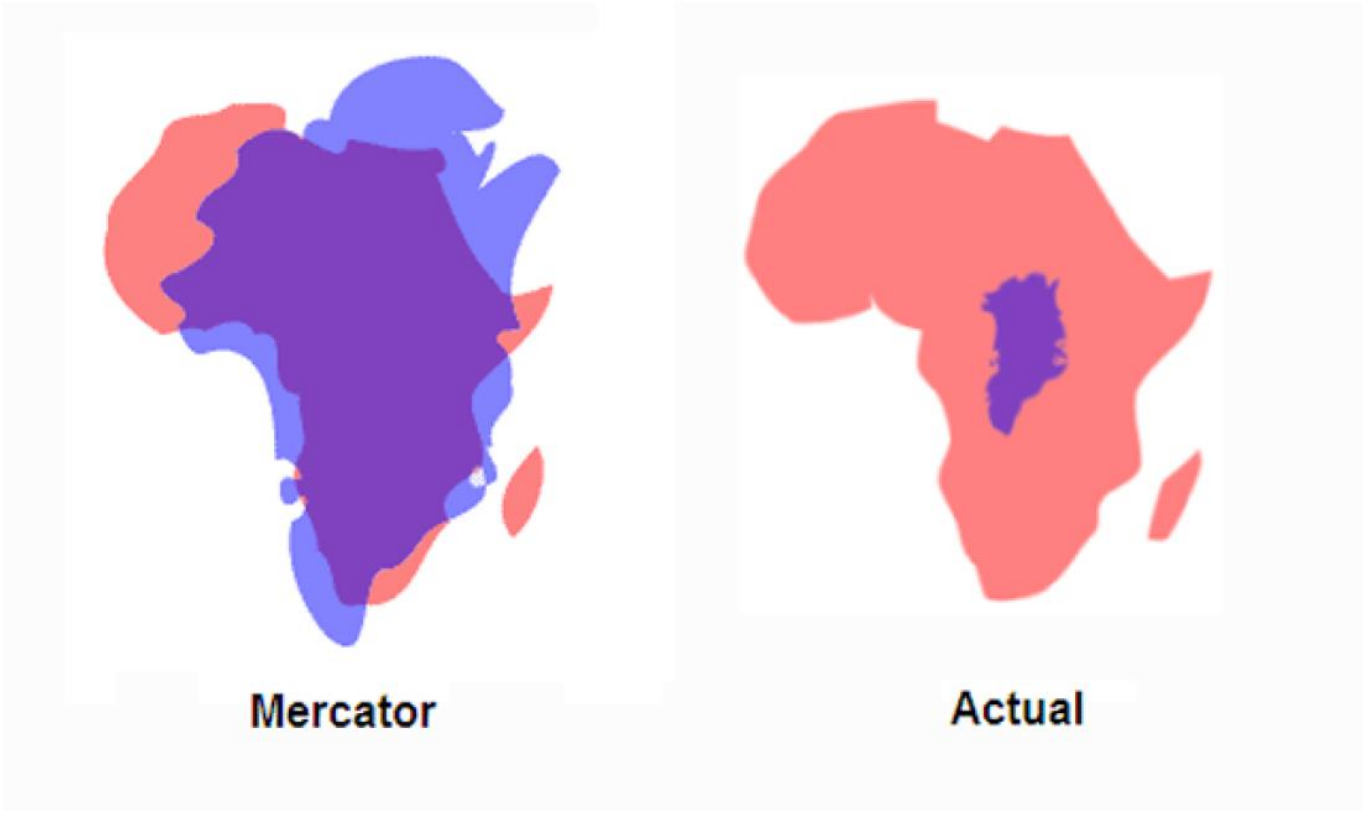
- 人类的认知是可以被欺骗的
- 操纵地图设计参数
  - 图例、颜色、区间等
- 通过特定的投影方法控制区域的面积
  - 使大的地区看起来更重要

世界地图

非洲

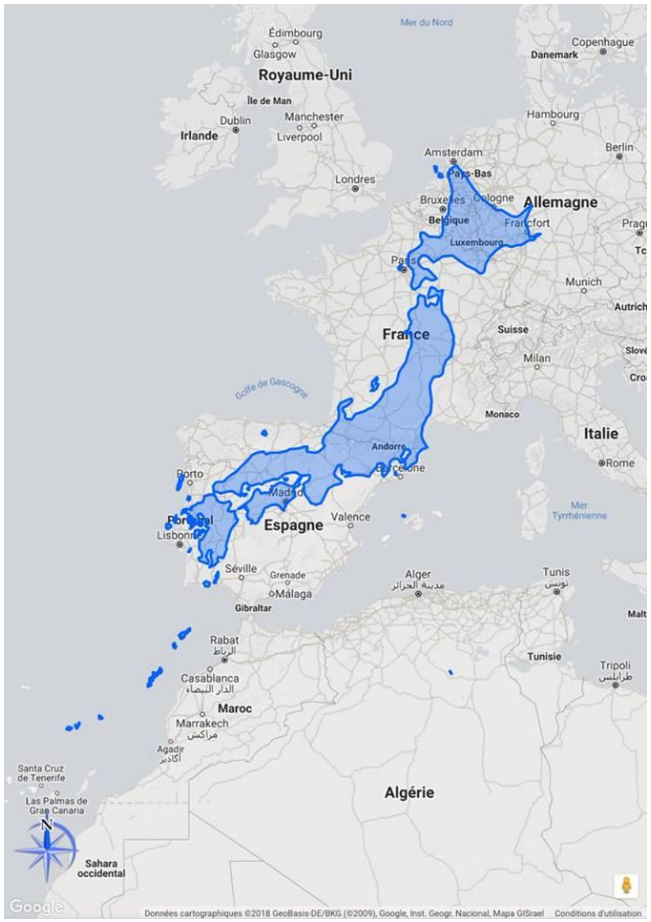

Africa vs Russia (actual size)
Africa
Area | 30.2 million km²
Russia
Area | 16.4 million km²
CNN

格陵兰岛



日本


JAPAN
Sapporo
Chicago
St. Louis
Washington, D.C.
Osaka
Tokyo
Dallas
North Atlantic Ocean
North Pacific Ocean
Gulf of Mexico
Durango
Mexico
0 500 Km
0 500 Mi


Royaume-Uni
Irlande
Allemagne
France
Espagne
Italie
Suisse
Monaco
Maroc
Algérie
Tunisie
Gibraltar
Mer du Nord
Mer Tyrrhénienne
Golfe de Gascogne
Sahara occidental
Google

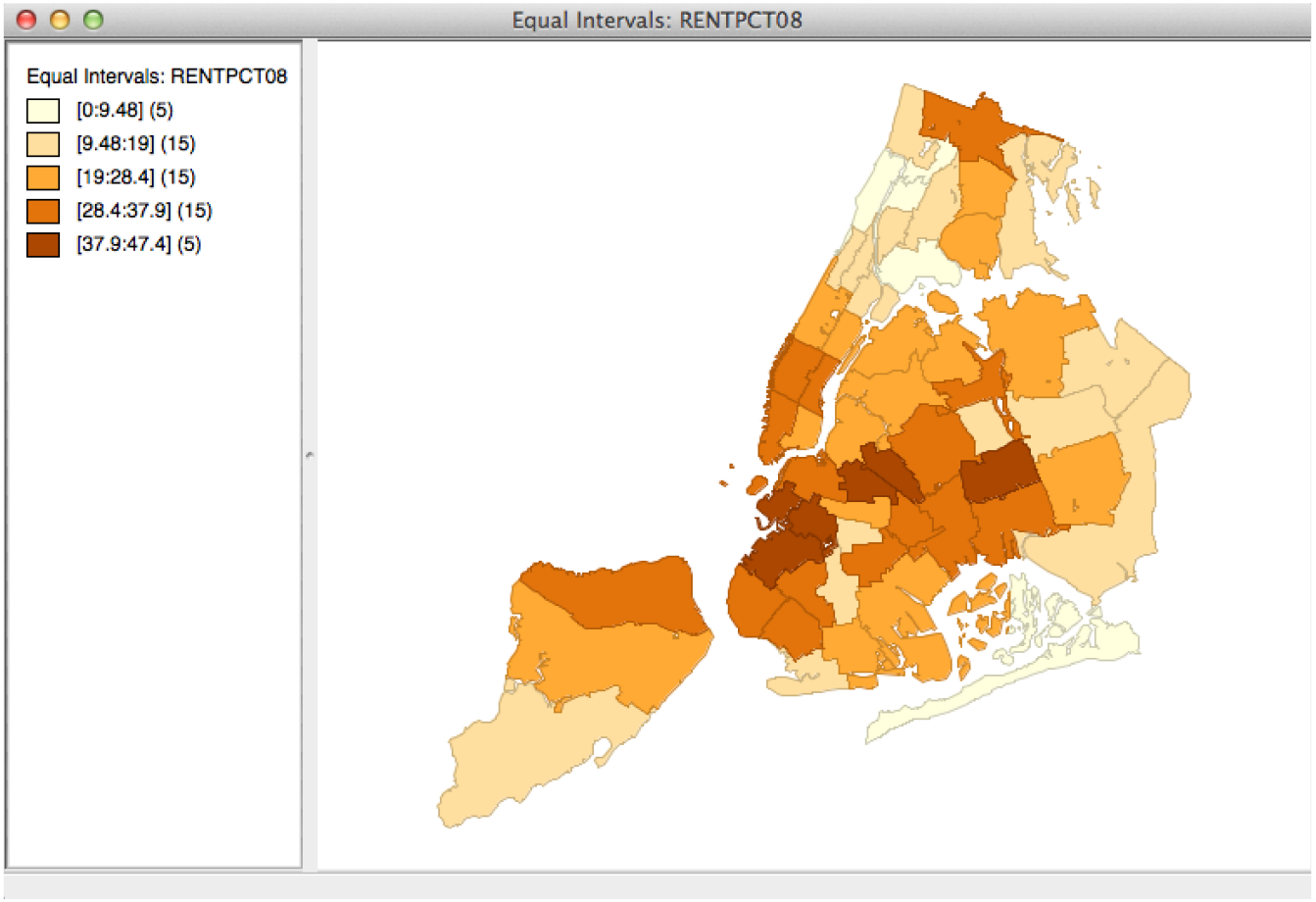
分级统计图

- 空间分布的可视化
- 目的类似于直方图，是对空间分布的描述
- 基本元素
  - 数据与坐标系（大地测量学）
  - 投影：形状、面积、距离、方向
  - 分类
  - 颜色
  - 图例

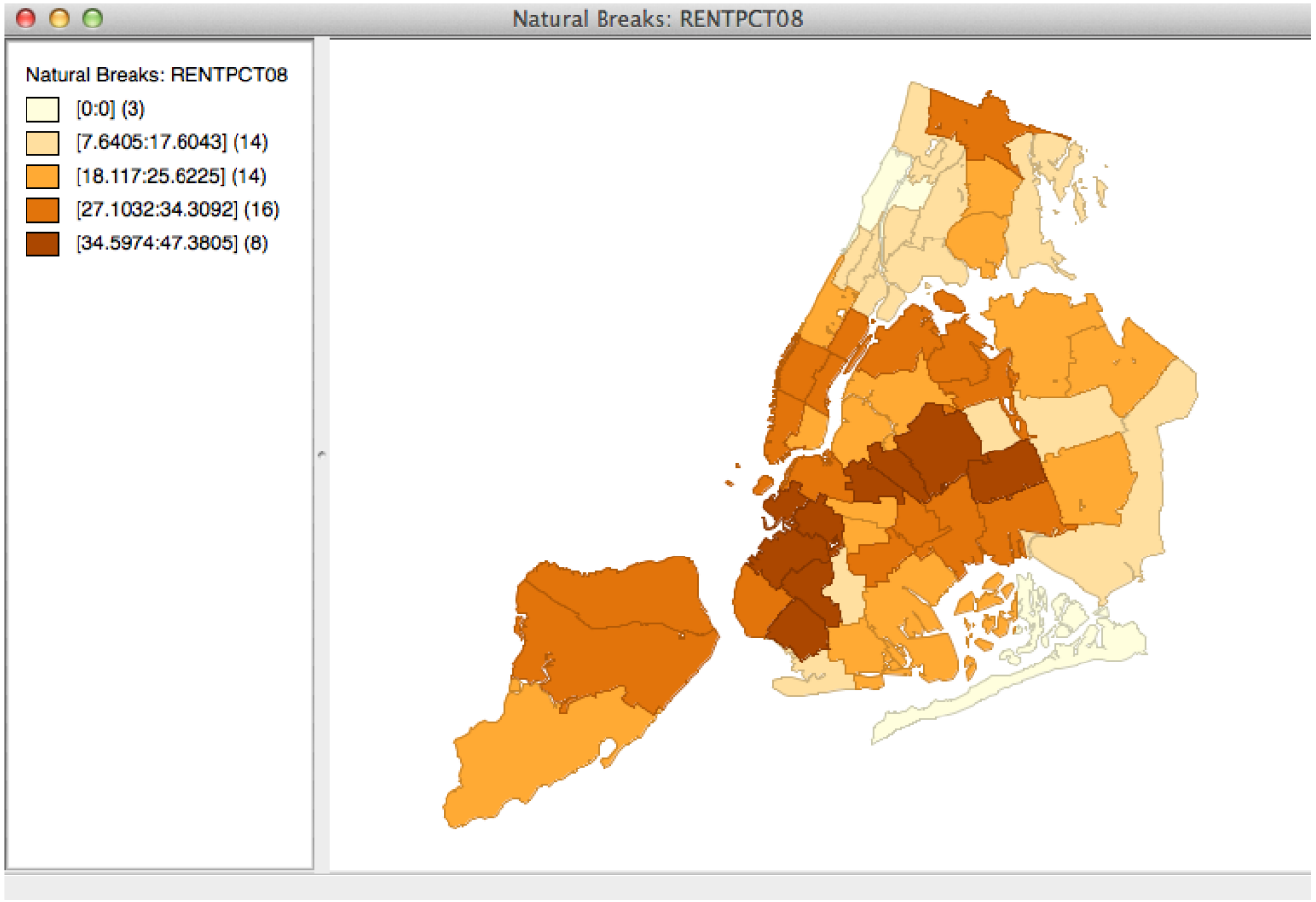
数值分类

- 数值：
  - 离散（选择区间、同区间内的数据颜色相同）
  - 连续（渐变色）
- 分类
  - 选择区间
  - 基于切割点
    - 等距、Jenks聚类（natural breaks，分组方差尽量小，组间方差尽量大）、人工定义
  - 基于统计指标
    - 分位数、标准差、极值

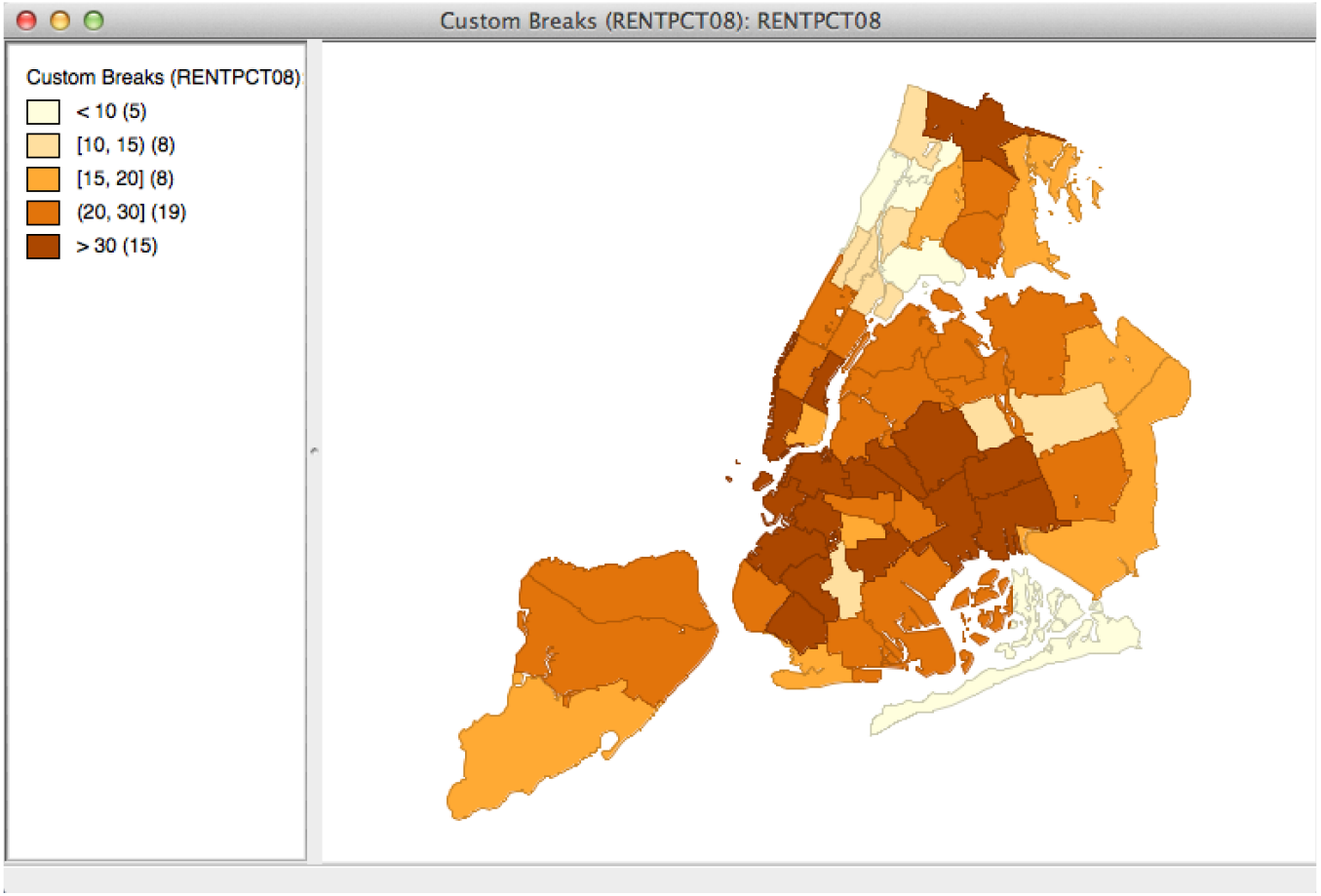
Equal Interval



Natural Breaks

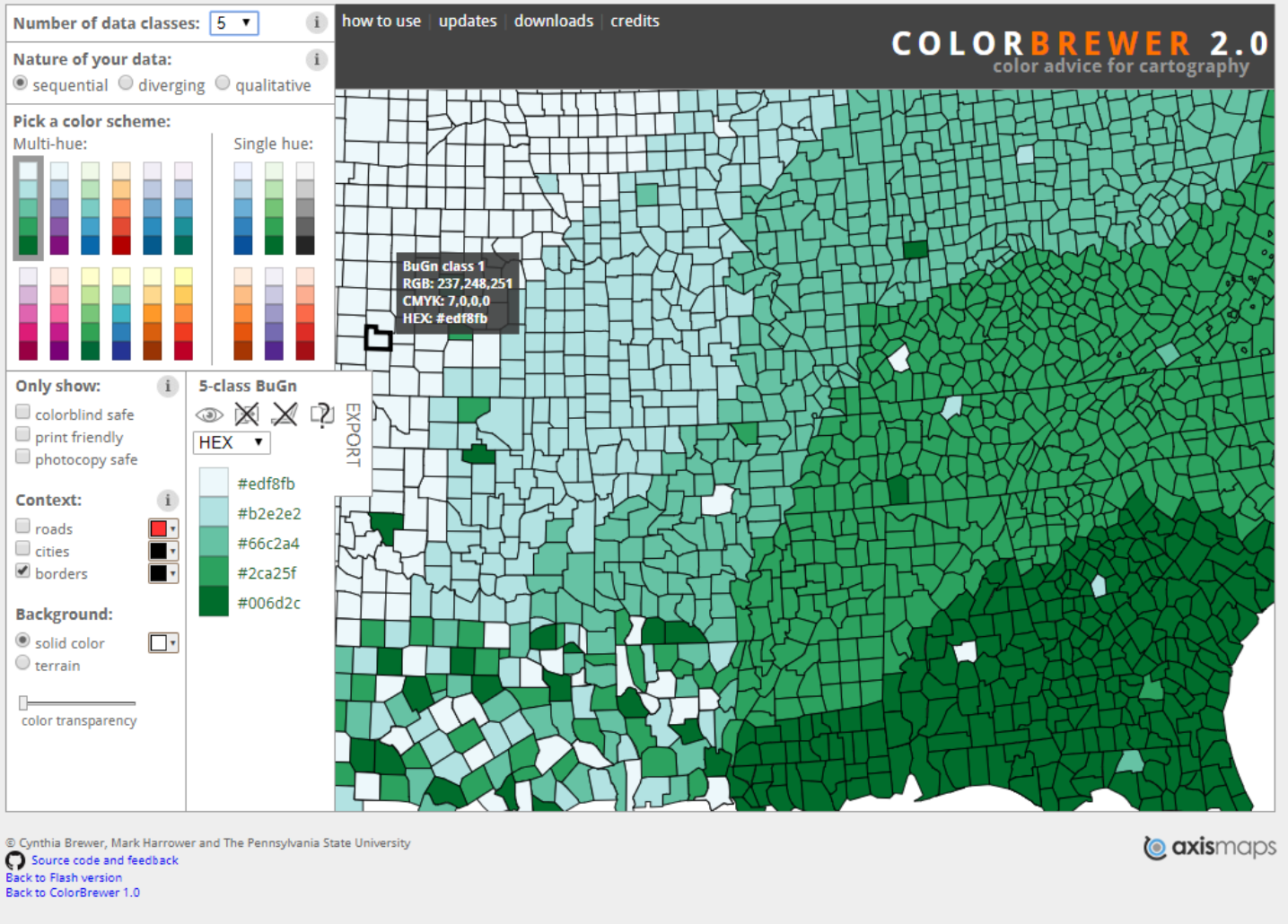


Custom Intervals

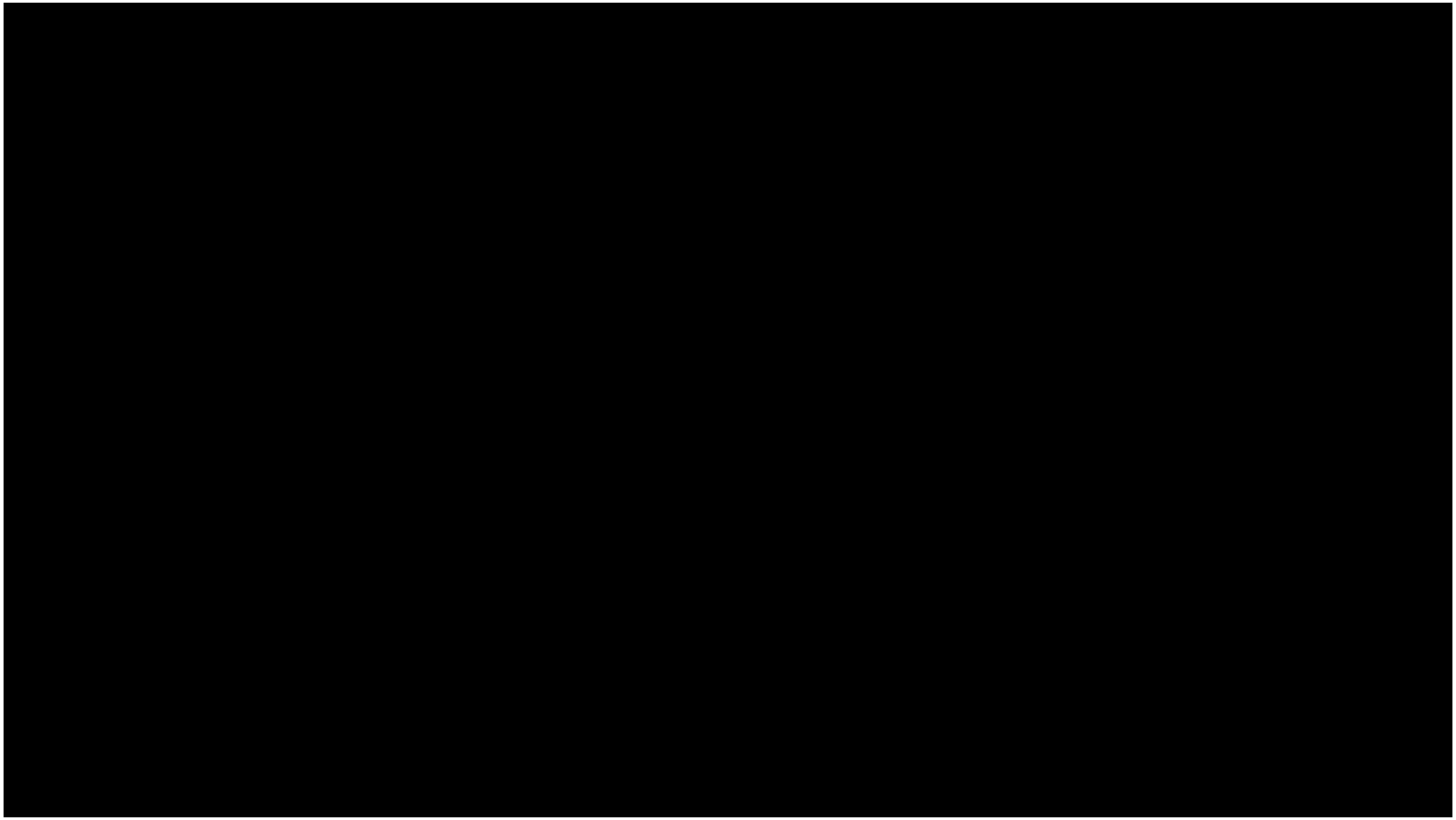


配色

配色（http://colorbrewer2.org）



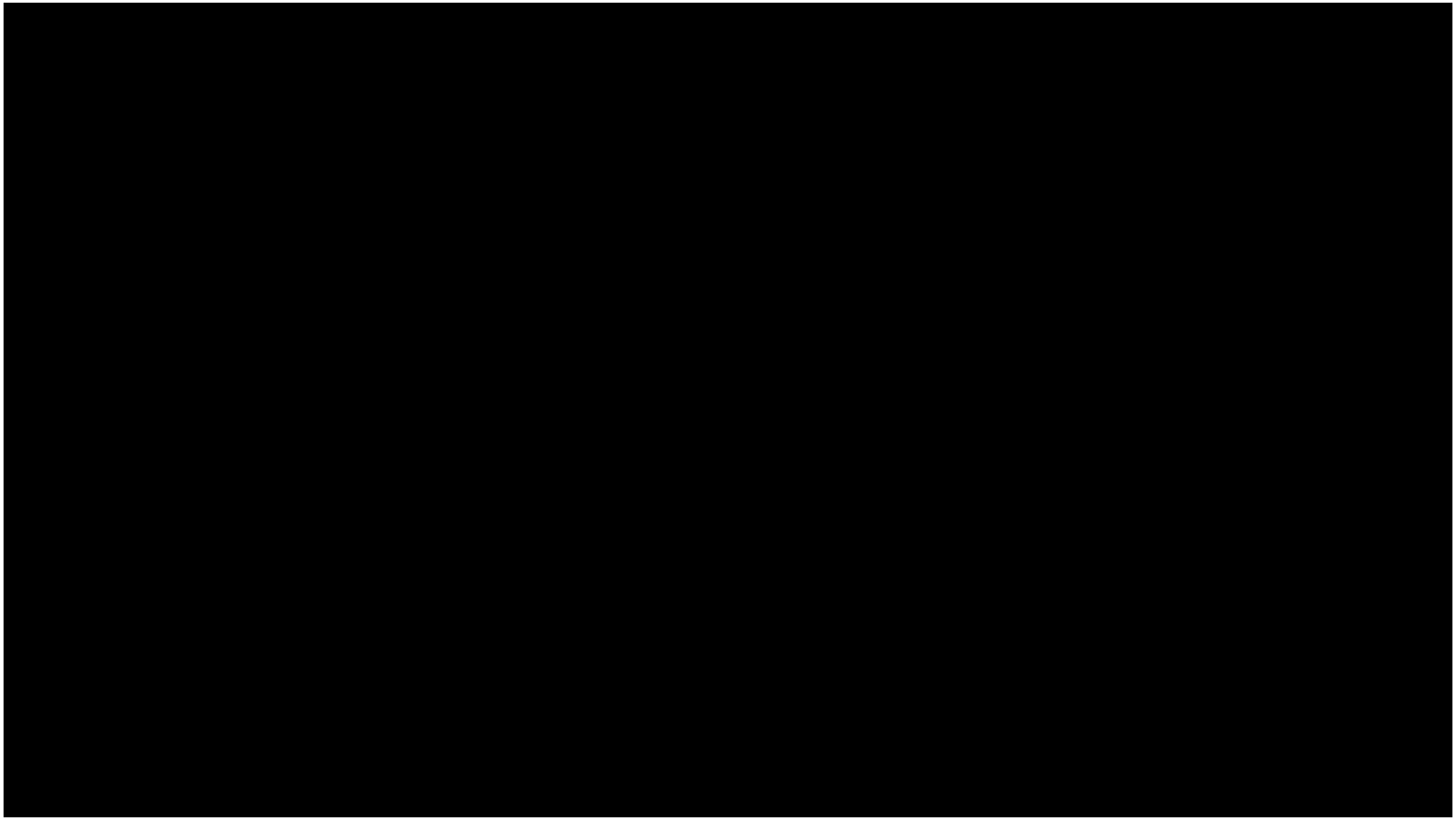
配色理论



可视化工具

- d3js.org
- deck.gl
- echartsjs.com

新加坡共享单车数据可视化



作业

- 基于波士顿的数据，选择一个变量
  - 设计空间权重矩阵
  - 探索其空间依赖关系
  - 建立空间回归模型
    - 判断应该具体选择哪个模型
- 实验报告（电子版）
  - 不超过10页
  - 默认页边距、小四（12号）、1.5倍行距
  - 宋体（英文、数字可以是Times New Roman或其他衬线字体）
- **截止时间：2022年04月22日上午9点59分**
- **提交邮箱：945441387@qq.com（吕叶婷）**

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

第七讲 结束